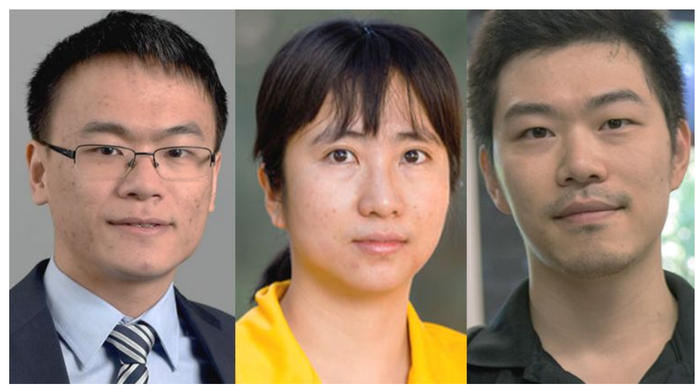
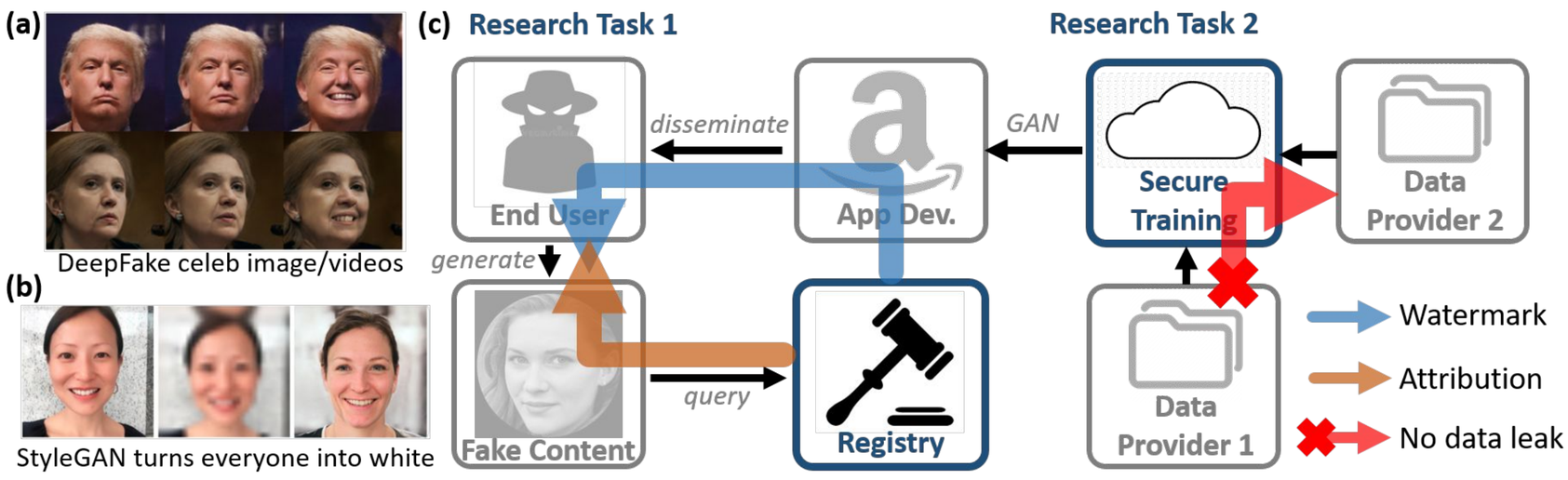
SaTC: Core: Small: Decentralized Attribution and Secure Training of Generative Models (# 2101052)



Yi Ren, Ni Trieu, 'YZ' Yezhou Yang | Arizona State University

https://chkimmmmm.github.io/SaTC_Decentralized/

- **Background:** Advances in generative models (e.g., for modeling media, behavioral, or scientific data) have demanded for investigation into security and privacy issues in their creation and dissemination.
- **Objective 1: <u>Model attribution</u>** - Generated contents should be attributed to their source models correctly.
- **Objective 2: <u>Secure training</u>** - Collaborative training should not expose proprietary data to other collaborators.

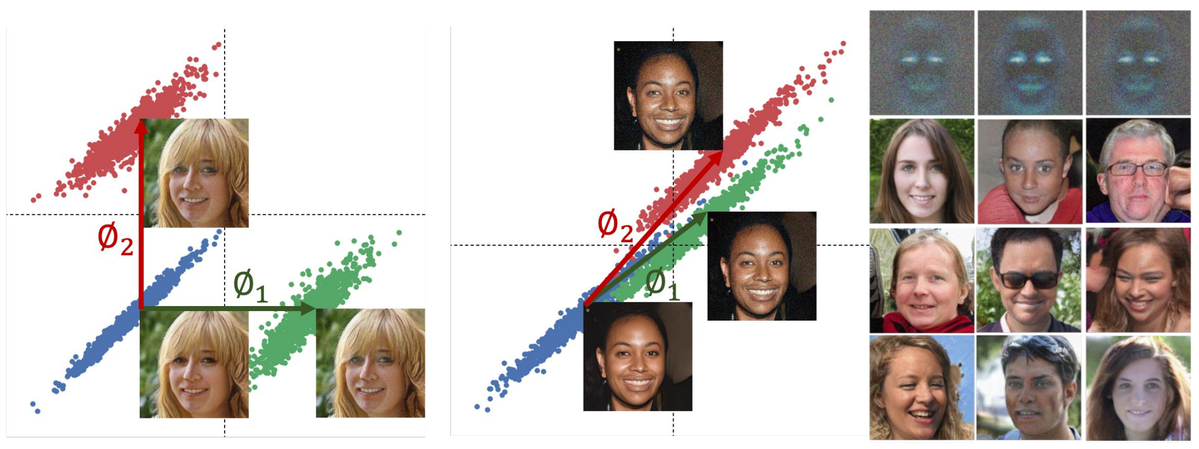


(a) Generative models for image and video synthesis have created socio-technical challenges. (b) Biases in generative models exist, motivating collaborative training and raising privacy concerns. (c) Project overview.

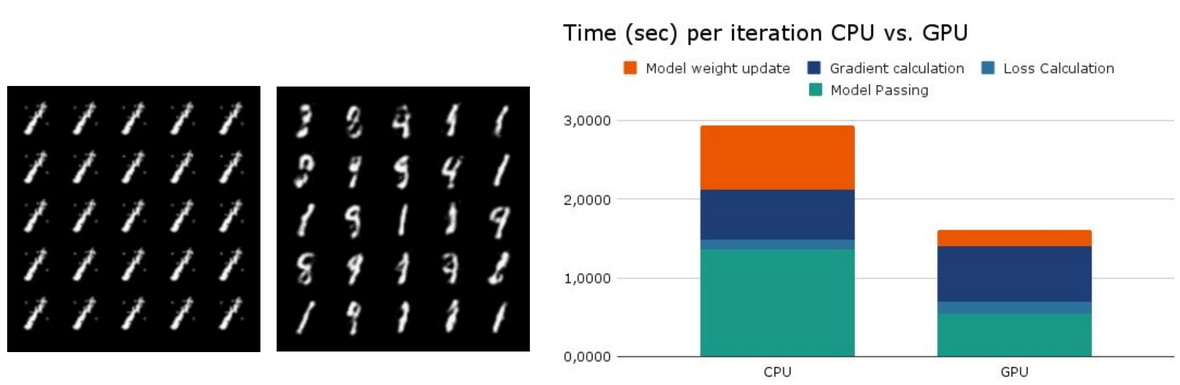
- **Challenge 1: Decentralized attribution** - No *provable* method for attributing user-specific contents w/o re-training a *centralized* classifier on ever growing user base.
- **Challenge 2: System tradeoff** - No *system-level* study on the trade-off among robust attribution (against adversarial de-attribution), generation quality, and the capacity of attributable models.
- **Challenge 3: Scalability** - Scalable secure training of generative models cannot be achieved due to limitations of secure multiplication of encrypted values and nonlinear activation/loss functions.

- **Contribution 1: Certification of decentralized attribution** - Sufficient conditions derived for content-space watermarking to achieve certifiable decentralized attribution, enabling effective watermarking for images and audios (ICLR2021, ICASSSP2022). Large robust attribution-generation quality tradeoff observed.
- **Contribution 2: Private join and compute (PJC)** - A generic solution that enables secure computation over multiple databases, and is applicable to privacy-preserving training of generative models (AsiaCrypt2021). Baseline implementation of secure GAN.

- **Scientific Impact 1: New computational solution to the open challenge of optimal packing** - Packing non-convex objects (model distributions) in a high-dim manifold via that in the latent space.
- **Scientific Impact 2:** New functionality – Private Join and Compute. New cryptographic primitive – private information retrieval (PIR) with default.



(Left, Middle) Watermarks w/ sufficient cosine-distance guarantees attributability. (Right) Three attributable watermarks on StyleGAN.



(Left, Middle) Generated digits w/o and w/ batch normalization from training on encrypted MNIST (Right) Time per secure training iteration on CPU vs. GPU.

- **Broader impact 1: Social -** Address threats, especially to underrepresented groups, from malicious personation (generative DeepFake) and biased data/model applications.
- **Broader impact 2: National security -** Secure training on scientific data (e.g., for manufacturing network).
- **Broader impact 3: Education and outreach -** Cross-disciplinary course materials on cyber-security, generative models, and optimization theory. High-school research engagement through ASU SCENE program.